Package 'gpart'

April 16, 2019

Title Human genome partitioning of dense sequencing data by identifying haplotype blocks

Version 1.0.3

Depends R (>= 3.5.0), grid, Homo.sapiens, TxDb.Hsapiens.UCSC.hg38.knownGene,

Description we provide a new SNP sequence partitioning method which partitions the whole SNP sequence based on not only LD block structures but also gene location information. The LD block construction for GPART is performed using Big-LD algorithm, with additional improvement from previous version reported in Kim et al.(2017). We also add a visualization tool to show the LD heatmap with the information of LD block boundaries and gene locations in the package.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

biocViews Software, Clustering

Imports igraph, biomaRt, Rcpp, data.table, OrganismDbi, AnnotationDbi, grDevices, stats, utils, GenomicRanges, IRanges

LinkingTo Rcpp

RoxygenNote 6.1.1

Suggests knitr, rmarkdown, BiocStyle, testthat

VignetteBuilder knitr

git_url https://git.bioconductor.org/packages/gpart

git_branch RELEASE_3_8

git_last_commit 7fdaae0

git_last_commit_date 2019-03-31

Date/Publication 2019-04-15

Author Sun Ah Kim [aut, cre, cph], Yun Joo Yoo [aut, cph]

Maintainer Sun Ah Kim <sunnyeesl@gmail.com>

R topics documented:

`BigLD` *Estimation of LD block regions*

Description

`BigLD` returns the estimation of LD block regions of given data.

Usage

```
BigLD(geno=NULL, SNPinfo=NULL,genofile=NULL, SNPinfofile=NULL,
cutByForce=NULL, LD=c("r2", "Dprime"), CLQcut=0.5,
clstgap=40000, CLQmode=c("density", "maximal"),
leng=200, subTaskSize=1500, MAFcut=0.05, appendRare=FALSE,
hrstType=c("near-nonhrst", "fast", "nonhrst"),
hrstParam=200, chrN=NULL, startbp=-Inf, endbp=Inf)
```

Arguments

`geno`	Data frame or matrix of additive genotype data, each column is additive genotype of each SNP.
`SNPinfo`	Data frame or matrix of SNPs information. 1st column is rsID and 2nd column is bp position.
`genofile`	Character constant; Genotype data file name (supporting format: .txt, .ped, .raw, .traw, .vcf).
`SNPinfofile`	Character constant; SNPinfo data file name (supporting format: .txt, .map).
`cutByForce`	Data frame; information of SNPs which are forced to be the last SNP LD block boundary.
`LD`	Character constant; LD measure to use, r2 or Dprime. Default "r2".
`CLQcut`	Numeric constant; a threshold for the LD measure value |r|, between 0 to 1. Default 0.5.
`clstgap`	Numeric constant; a threshold of physical distance (bp) between two consecutive SNPs which do not belong to the same clique, i.e., if a physical distance between two consecutive SNPs in a clique greater than `clstgap`, then the algorithm split the cliques satisfying each clique do not contain such consecutive SNPs. Default 40000.

`CLQmode`	Character constant; the way to give priority among detected cliques. if `CLQmode = "density"` then the algorithm gives priority to the clique of largest value of $(Numberof SNPs)/(rangeof cliqu$ else if `CLQmode = "maximal"`, then the algorithm gives priority to the largest clique. The default is "density".
`leng`	Numeric constant; the number of SNPs in a preceding and a following region of each sub-region boundary, every SNP in a preceding and every SNP in a following region need to be in weak LD. Default 200.
`subTaskSize`	Numeric constant; upper bound of the number of SNPs in a one-take sub-region. Default 1500.
`MAFcut`	Numeric constant; the MAF threshold. Default 0.05.
`appendRare`	logical; if TRUE, the function append rare variants with MAF<MAFcut to the constructed blocks.
`hrstType`	Character constant; heuristic methods. If you want to do not use heuristic algorithm, set `hrstType = "nonhrst"`. If you want to use heuristic algorithm suggested in Kim et al.,(2017), set `hrstType = "fast"`. That algorithm is fastest heuristic algorithm and suitable when your memory capacity is not greater than 8GB. If you want to obtain the results similar to the that of non-heuristic algorithm, set `hrstType = "near-nonhrst"`.
`hrstParam`	Numeric constant; parameter for heuristic algorithm "near-nonhrst". Default is `200`. It is recommended that you set the parameter to greater than 150.
`chrN`	Numeric(or Character) constant (or vector); chromosome number to use.
`startbp`	Numeric constant; starting bp position of the `chrN`. Default -Inf.
`endbp`	Numeric constant; last bp position of the `chrN`. Default Inf.

Value

A data frame of block estimation result. Each row of data frame shows the starting SNP and end SNP of each estimated LD block.

Author(s)

Sun-Ah Kim <sunny03@snu.ac.kr>, Yun Joo Yoo <yyoo@snu.ac.kr>

See Also

`CLQD`, `LDblockHeatmap`

Examples

```
data(geno)
data(SNPinfo)
BigLD(geno[,1:100], SNPinfo[1:100,])

## Not run:
BigLD(geno, SNPinfo, LD = "Dprime")
BigLD(geno, SNPinfo, CLQcut = 0.5, clstgap = 40000, leng = 200, subTaskSize = 1500)

## End(Not run)
```

CLQD *partitioning into cliques*

Description

CLQD partitioning the given data into subgroups that contain SNPs which are highly correlated.

Usage

```
CLQD(geno, SNPinfo, CLQcut=0.5, clstgap=40000,
hrstType=c("near-nonhrst", "fast", "nonhrst"), hrstParam=200,
CLQmode=c("density", "maximal"), LD=c("r2", "Dprime"))
```

Arguments

geno	Data frame or matrix of additive genotype data, each column is additive genotype of each SNP. (Use data of non-monomorphic SNPs)
SNPinfo	Data frame or matrix of SNPs information. 1st column is rsID and 2nd column is bp position.
CLQcut	Numeric constant; a threshold for the LD measure value |r|, between 0 to 1. Default 0.5.
clstgap	Numeric constant; a threshold of physical distance (bp) between two consecutive SNPs which do not belong to the same clique, i.e., if a physical distance between two consecutive SNPs in a clique greater than clstgap, then the algorithm split the cliques satisfying each clique do not contain such consecutive SNPs. Default 40000.
hrstType	Character constant; heuristic methods. If you want to do not use heuristic algorithm, set hrstType = "nonhrst". If you want to use heuristic algorithm suggested in Kim et al.,(2017), set hrstType = "fast". That algorithm is fastest heuristic algorithm and suitable when your memory capacity is not greater than 8GB. If you want to obtain the results similar to the that of non-heuristic algorithm, set hrstType = "near-nonhrst".
hrstParam	Numeric constant; parameter for heuristic algorithm "near-nonhrst". Default is 200. It is recommended that you set the parameter to greater than 150.
CLQmode	Character constant; the way to give priority among detected cliques. if CLQmode = "density" then the algorithm gives priority to the clique of largest value of $(NumberofSNPs)/(rangeofcliqu$ else if CLQmode = "maximal", then the algorithm gives priority to the largest clique. The default is "density".
LD	Character constant; LD measure to use, "r2" or "Dprime". Default "r2".

Value

A vector of cluster numbers of all SNPs (NA represents singleton cluster).

Author(s)

Sun-Ah Kim <sunny03@snu.ac.kr>, Yun Joo Yoo <yyoo@snu.ac.kr>

See Also

BigLD

Examples

```
data(geno)
data(SNPinfo)
CLQD(geno=geno[,1:100],SNPinfo=SNPinfo[100,])
CLQD(geno=geno[,1:100],SNPinfo=SNPinfo[100,], CLQmode = 'maximal')
CLQD(geno=geno[,1:100],SNPinfo=SNPinfo[100,], LD='Dprime')
```

---

convert2GRange *convert output of BigLD/GPART to "GRangesList" object*

Description

convert2GRange convert a BigLD or GPART output to a data of GRangesList object.

Usage

```
convert2GRange(blockresult)
```

Arguments

blockresult	BigLD or GPART output

Value

GRangeList object including the BigLD or GPART output

Author(s)

Sun-Ah Kim <sunny03@snu.ac.kr>, Yun Joo Yoo <yyoo@snu.ac.kr>

Examples

```
testBigLD <- BigLD(geno=geno[,1:100], SNPinfo=SNPinfo[1:100,])
testBigLD_grange <- convert2GRange(testBigLD)
```

geneinfo *gene information data*

Description

This data set gives gene information in chromosome 21.

Usage

```
data(geneinfo)
```

Format

A data frame with 736 rows and 4 columns for genename, chromosome name, start bp and end bp of each gene.

Source

<http://grch37.ensembl.org/index.html>

References

Zerbino, Daniel R., et al. "Ensembl 2018." *Nucleic acids research 46.D1* (2017): D754-D761.

geno *genotype data*

Description

This data set gives genotype data that are subset of 1000 Genomes Project phase1 release 3 genotype data with 286 individuals from JPT, CHB and CHS populations (1000 Genomes Project Consortium, 2012). The dataset contains 9000 SNPs in chromosome 21

Usage

```
data(geno)
```

Format

A data frame with 286 rows and 9000 columns

Source

1000 genomes project Phase 1 dataset <http://www.internationalgenome.org/>

References

1000 Genomes Project Consortium. "An integrated map of genetic variation from 1,092 human genomes." *Nature* 491.7422 (2012): 56.

GPART *Partitioning genodata based on the result obtained by using Big-LD and gene region information.*

Description

GPART partition the given genodata using the result obtained by Big-LD and gene region information. The algorithm partition the whole sequence into sub sequences of which size do not exceed the given threshold.

Usage

```
GPART(geno=NULL, SNPinfo=NULL, geneinfo=NULL, genofile=NULL,
SNPinfofile=NULL, geneinfofile=NULL, geneDB = c("ensembl","ucsc"),
assembly = c("GRCh38", "GRCh37"), geneid = "hgnc_symbol",ensbversion = NULL,
chrN=NULL, startbp=-Inf, endbp=Inf, BigLDresult=NULL, minsize=4, maxsize=50,
LD=c("r2", "Dprime"), CLQcut=0.5, CLQmode=c("density", "maximal"), MAFcut = 0.05,
GPARTmode=c("geneBased", "LDblockBased"),
Blockbasedmode=c("onlyBlocks", "useGeneRegions"))
```

Arguments

geno	Data frame or matrix of additive genotype data, each column is additive genotype of each SNP.
SNPinfo	Data frame or matrix of SNPs information. 1st column is rsID and 2nd column is bp position.
geneinfo	Data frame or matrix of Gene info data. (1st col : Genename, 2nd col : chromosome, 3rd col : start bp, 4th col : end bp)
genofile	Character constant; Genotype data file name (supporting format: .txt, .ped, .raw, .traw, .vcf).
SNPinfofile	Character constant; SNPinfo data file name (supporting format: .txt, .map).
geneinfofile	A Character constant; file containing the gene information (1st col : Genename, 2nd col : chromosome, 3rd col : start bp, 4th col : end bp)
geneDB	A Character constant; database type for gene information. Set "ensembl" to get gene info from "Ensembl", or set "ucsc" to get gene info from "UCSC genome browser" (See package "biomaRt" or package "homo.sapiens"/"TxDb.Hsapiens.UCSC.hg38.knownG "TxDb.Hsapiens.UCSC.hg19.knownGene" for details.)
assembly	A character constant; set "GRCh37" for GRCh37, or set "GRCh38" for GRCh38
geneid	A character constant; When you use the gene information by geneDB. specity the symbol for gene name to use. default is "hgnc_symbol". (eg. 'ensembl_gene_id' for geneDB = "ensembl", "ENTREZID"/"ENSEMBL"/"REFSEQ"/"TXNAME" for geneDB="ucsc". See package 'biomaRt' or package 'Homo.sapiens' for details)
ensbversion	a integer constant; you can set the release version of ensembl when you use the gene information by using geneDB='emsembl' and assembly='GRCh38'
chrN	Numeric(or Character) constant (or vector); chromosome number to use. If NULL(default), we use all chromosome.
startbp	Numeric constant; starting bp position of the chrN. Default -Inf.

endbp	Numeric constant; last bp position of the chrN. Default Inf.
BigLDresult	Data frame; a result obtained by BigLD function. If NULL(default), the GPART function first excute BigLD function to obtain LD blocks estimation result.
minsize	Numeric constant; the lower bound of number of SNPs in a partition.
maxsize	Numeric constant; the upper bound of number of SNPs in a partition.
LD	Character constant; LD measure to use, r2 or Dprime.
CLQcut	Numeric constant; threshold for the correlation value |r|, between 0 to 1.
CLQmode	Character constant; the way to give priority among detected cliques. if CLQmode = "density" then the algorithm gives priority to the clique of largest value of $(NumberofSNPs)/(rangeofcliqu$ else if CLQmode = "maximal", then the algorithm gives priority to the largest clique. The default is "density".
MAFcut	Numeric constant; the MAF threshold. Default 0.05.
GPARTmode	Character constant; GPART algorithm methods to use, "geneBased" or "LDblockBased". Default is ??geneBased??
Blockbasedmode	Character constant; When you set GPARTmode = "LDblockBased", specify LDblock based method as "onlyBlocks"("LDblock based only" algorithm) or "useGeneRegions"(LDblock based and also use gene info algorithm).

Value

GPART returns data frame which contains 9 information of each partition (chromosome, index number of the first SNP and last SNP, rsID of the first SNP and last SNP, basepair position of the first SNP and last SNP, blocksize, Name of a block)

Author(s)

Sun Ah Kim <sunny03@snu.ac.kr>, Yun Joo Yoo <yyoo@snu.ac.kr>

See Also

BigLD

Examples

```
data(geno)
data(SNPinfo)
data(geneinfo)
GPART(geno=geno[,1:100], SNPinfo=SNPinfo[1:100,], geneinfo=geneinfo)
```

LDblockHeatmap *visualization of LD block structure*

Description

LDblockHeatmap visualize the LD structure or LD block results of the inputed data. LDblockHeatmap <- function(geno=NULL, SNPinfo=NULL, genofile=NULL, SNPinfofile=NULL,geneinfo=NULL, geneinfofile = NULL, geneDB = c("ensembl","ucsc","file"), assembly = c("GRCh38", "GRCh37"), geneid = "hgnc_symbol", ensbversion = NULL, chrN=NULL,startbp=-Inf, endbp=Inf, blockresult=NULL, blocktype=c("bigld", "gpart"), minsize=4, maxsize=50, LD=c("r2", "Dprime", "Dpstr"), MAFcut=0.05,CLQcut=0.5, CLQmode=c("density", "maximal"), CLQshow=FALSE, type=c("png", "tif"), filename="heatmap", res=300, onlyHeatmap=FALSE)

Usage

```
LDblockHeatmap(geno = NULL, SNPinfo = NULL, genofile = NULL,
  SNPinfofile = NULL, geneinfo = NULL, geneinfofile = NULL,
  geneDB = c("ensembl", "ucsc", "file"), assembly = c("GRCh38",
  "GRCh37"), geneid = "hgnc_symbol", ensbversion = NULL,
  geneshow = TRUE, chrN = NULL, startbp = -Inf, endbp = Inf,
  blockresult = NULL, blocktype = c("bigld", "gpart"), minsize = 4,
  maxsize = 50, LD = c("r2", "Dprime", "Dp-str"), MAFcut = 0.05,
  CLQcut = 0.5, CLQmode = c("density", "maximal"), CLQshow = FALSE,
  type = c("png", "tif"), filename = "heatmap", res = 300,
  onlyHeatmap = FALSE)
```

Arguments

geno	Data frame or matrix of additive genotype data, each column is additive genotype of each SNP.
SNPinfo	Data frame or matrix of SNPs information. 1st column is rsID and 2nd column is bp position.
genofile	Character constant; Genotype data file name (supporting format: .txt, .ped, .raw, .traw, .vcf).
SNPinfofile	Character constant; SNPinfo data file name (supporting format: .txt, .map).
geneinfo	Data frame or matrix of Gene info data. (1st col : Genename, 2nd col : chromosome, 3rd col : start bp, 4th col : end bp)
geneinfofile	A Character constant; file containing the gene information (1st col : Genename, 2nd col : chromosome, 3rd col : start bp, 4th col : end bp)
geneDB	A Character constant; database type for gene information. Set "ensembl" to get gene info from "Ensembl", or set "ucsc" to get gene info from "UCSC genome browser" (See package "biomaRt" or package "homo.sapiens"/"TxDb.Hsapiens.UCSC.hg38.knownG "TxDb.Hsapiens.UCSC.hg19.knownGene" for details.)
assembly	A character constant; set "GRCh37" for GRCh37, or set "GRCh38" for GRCh38
geneid	A character constant; When you use the gene information by geneDB. specity the symbol for gene name to use. default is "hgnc_symbol". (eg. 'ensembl_gene_id' for geneDB = "ensembl", "ENTREZID"/"ENSEMBL"/"REFSEQ"/"TXNAME" for geneDB="ucsc". See package 'biomaRt' or package 'Homo.sapiens' for details)
ensbversion	a integer constant; you can set the release version of ensembl when you use the gene information by using geneDB='emsembl' and assembly='GRCh38'
geneshow	logical; do not show the gene information if geneshow=FALSE. Default is geneshow=TRUE
chrN	Numeric(or Character) constant ; chromosome number to use. If the data contains more than one chromosome, you need to specify the chromosome to show.
startbp	Numeric constant; starting bp position of the chrN.
endbp	Numeric constant; last bp position of the chrN.
blockresult	Data frame; a result obtained by BigLD function or GPART. If NULL(default), the function first excute BigLD orGPART to obtain block estimation result denpending on the blocktype.
blocktype	Character constant; "bigld" for Big-LD or "gpart" for GPART. Default is "gpart".
minsize	Integer constant; when blockresult=NULL, blocktype="gpart" specify the threshold for minsize of a block obtained by GPART

maxsize	Integer constant; when blockresult=NULL, blocktype="gpart" specify the threshold for maxsize of a block obtained by GPART
LD	Character constant; LD measure to use, "r2" or "Dprime" or "Dp-str". LD measures for LD heatmap (and BigLD execution when BigLDresult=NULL). When LD = "Dp-str", heatmap shows only two cases, "weak LD or not-informative" and "strong LD". When LD= Dprime heatmap shows the estimated D' measures.
MAFcut	Numeric constant; MAF threshold of SNPs to use. Default 0.05
CLQcut	Numeric constant; threshold for the correlation value |r|, between 0 to 1. Default 0.5.
CLQmode	Character constant; the way to give priority among detected cliques. if CLQmode = "density" then the algorithm gives priority to the clique of largest value of $(Numberof SNPs)/(rangeof cliqu$ else if CLQmode = "maximal", then the algorithm gives priority to the largest clique. Default "density".
CLQshow	logical; Show the LD bin structures, i.e. CLQ results, in each LD blocks. Notice that if the region to show includes more than 200 SNPs, the function do now show LD bin structures automatically. Default false.
type	Character constant; file format of output image file. set png for PNG format file, or tif for TIFF format file.
filename	Character constant; filename of output image file. Default "heatmap".
res	Numeric constant; resolution of image. Default 300.
onlyHeatmap	logical; show the LD heatmap without the LD block boundaries. Default false.

Value

GPART returns data frame which contains 9 information of each partition (chromosome, index number of the first SNP and last SNP, rsID of the first SNP and last SNP, basepair position of the first SNP and last SNP, blocksize, Name of a block)

Author(s)

Sun-Ah Kim <sunny03@snu.ac.kr>, Yun Joo Yoo <yyoo@snu.ac.kr>

See Also

BigLD

Examples

```
LDblockHeatmap(geno=geno[,1:100], SNPinfo=SNPinfo[1:100,], geneinfo=geneinfo,
filename="chr21Heatmap")
```

SNPinfo *SNP information data*

Description

This data set gives information data of SNPs in geno The dataset contains chromosome name, rsID and bp information of the 9000 SNPs in geno

Usage

```
data(SNPinfo)
```

Format

A data frame with 9000 rows and 3 columns for chromosome name, rsID and bp

Source

1000 genomes project Phase 1 dataset <http://www.internationalgenome.org/>

References

1000 Genomes Project Consortium. "An integrated map of genetic variation from 1,092 human genomes." *Nature* 491.7422 (2012): 56.

@keywords datasets

Index